WARNING: UNFINISHED DRAFT

THE COGAFF PROJECT

Papers and presentations on affect, in the Birmingham Cognition and Affect Project started here in 1991, building on earlier work at Sussex University.

Aaron Sloman
http://www.cs.bham.ac.uk/~axs/
School of Computer Science, University of Birmingham

This document is available in two formats:
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/emotions-affect.html
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/emotions-affect.pdf

NOTE

I've always thought that great novelists know more about emotions, and other affective phenomena than psychologists and neuroscientists ever will.

(DRAFT: Liable to change)
Last updated: 11 Apr 2024
More Recent Items, below

A major new web site is now under development here, subsuming many of the issues discussed or mentioned below:
https://www.cs.bham.ac.uk/research/projects/cogaff/misc/metamorphosis.html
This is includes references to the largest and most complex collection of related but varied ideas I have ever assembled. I don't know whether anyone else has noticed and written about all those connections, although many others have addressed significant subsets -- different subsets!

Other (mostly much older) papers and presentations more concerned with non-affective aspects, e.g. perception, reasoning, learning are included in:
http://www.cs.bham.ac.uk/research/projects/cogaff/
with talks/presentations here
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/

What is this?

This (still incomplete) document lists some of the papers written at the University of Birmingham (1991 onwards), mainly within the Cognition and Affect (CogAff) project, concerned with a general framework for combining affective mechanisms (involved in wants, hopes, fears, likes, dislikes, emotions, moods, evaluations, ...etc...etc.) with cognitive phenomena (e.g. perceiving, acting, planning, learning, predicting, explaining, learning, hypothesizing, ...etc...etc.)

From its early days the project was committed to a theoretical-comparative approach, i.e. studying not only phenomena that occur in humans, or in humans and other animals, but studying the whole space of possible designs (especially possible information processing architectures) for systems capable of having informational states and processes (involving desires, preferences, values, moods, beliefs, skills, knowledge, uncertainty, etc.) as well as having physical states and processes (including body form, actions available, sensors, motors, internal physiological systems, size, shape, weight, strength, changing physical needs and stored resources, etc.), in various environments (e.g. under water, on land, while flying) on flat terrain, in various types of non-flat environment (e.g. rocky, snowy, icy, muddy mountain slopes, etc.), with various ranges of temperature, resource availability, threats and dangers, etc.

This approach contrasts strongly both with shallow theories of embodied, enactive, expressive ("skin deep"??) aspects of affective states and processes, and utility-based theories of motivation, and also with informationally-restricted theories, e.g. assuming that all information and information processing is logical, or numerically measurable, or symbolic, or probability-based, or restricted in some other way (as happened at various stages in the history of AI theories of emotion).

The emphasis on *designs* and what they can and cannot do also contrasts with a focus on classifying and correlating measurable, or observable or introspectable or physiological states of humans or other species. A *design* is something deeper and more abstract, and may have multiple different sub-types of instantiation of the design. Compare the space of possible designs for utterances in the English language, or some other language -- which may have a different space -- or the space of types of communication modality: spoken, written, signed, signalled (e.g. using semaphore), etc.

The CogAff approach to the study of the space of possible minds, and possible mind-based states and processes, aims to bring about a change in science that is partly similar to how the periodic table of the elements initiated deep changes in chemistry. Unfortunately, most researchers on cognition and affect are not educated with the required attitudes, concepts, knowledge and skills, e.g. abilities to design, build, test, and debug working models.

The list below includes some relevant earlier papers from the time (pre-1991) that the project was based at Sussex University (1962-91).

Thanks to help from Dean Petters, the following includes a new architecture-schema diagram, showing overlaps between input, output, and "central" components of the architecture:
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/vm-functionalism.html
(also PDF)
N.B. There is no intended claim that all instances of the design must include instances of all the types of architectural feature: this is intended to be more like the specification grammatical structures for sentences, clauses, phrases, or larger linguistic products (e.g. stories) available in a

particular language. However, it is arguable that the variety of types of design specified here is not rich enough to accommodate all relevant types of design for systems capable of emotions or other affective states.

Who knows what?
Great poets, playwrights, novelists and composers often have a much deeper understanding of varieties of affect than (current) philosophers, psychologists, neuroscientists and cognitive modellers. But their deep understanding is implicit and usually only indirectly articulated, e.g. in plot construction, dialogue construction, thought-streams, musical compositions, etc.

One goal for AI is to find ways to make that implicit knowledge explicit and demonstrate the implications by building a succession of increasingly realistic, increasingly complex, working systems. But "working" does not merely mean showing behaviours (including linguistic behaviours) thought to correspond to various cognitive and affective states and processes. There must also be the right kinds in internal/invisible information processing including forms of reasoning, clashes of motivation, resolution of conflicts, growth and modification of attitudes and values, etc.

A full model should include "genetic time-bombs" i.e. potential at various late stages of development to produce new motives, values, preferences, abilities, etc. In humans the motivational (and consequential emotional) changes at puberty are obvious examples. But genetic time-bombs may have even longer fuses concerned with how to use a large volume of acquired knowledge, skills, experience, etc. after enough time has been spent on acquisition. (The corresponding mechanisms in humans seem to be highly erratic, and often over-ridden by self-interested motivation.)

NOTE:
Beware of arguments purporting to prove that "AI systems (including robots) can’t do X" by proving that "Computers can’t do X".

They are as valid as arguments showing that "Molecules can’t do Y (e.g. have emotions, or discover geometric theorems), therefore objects composed of molecules can’t do Y".

Turing machines, are irrelevant to AI for reasons explained in this paper:
http://www.cs.bham.ac.uk/research/cogaff/00-02.html#77

On the other hand complex systems composed of large numbers of interconnected digital computers, sensors and motors, in a machine located in a complex, changing, partly unpredictable environment, are another matter. (As H.A. Simon pointed out in "Motivational and emotional controls of cognition", 1967.)

The deep, still unanswered, question is: what sort of (self-extending) information-processing architecture could replicate the required functionality in future machines?

The answer may be related to the question whether sub-neuronal molecular computations are essential to biological competences of poets, playwrights, squirrels defeating squirrel-proof bird-feeders, and ancient mathematicians.

Clearly some motives e.g. hunger, are based on molecular processes. Perhaps far more aspects of mentality are than we realise. I have raised that question in connection with ancient spatial reasoning abilities underlying discoveries in geometry and topology, here:
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/kant-maths.html
Key Aspects of Immanuel Kant's Philosophy of Mathematics
(That's a companion-piece to a discussion of Turing's distinction in his PhD thesis between mathematical intuition and mathematical ingenuity: he suggested that computers could replicate the latter but not the former.
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/turing-intuition.html

Some incomplete remarks on requirements for types of computer capable of replicating spatial reasoning in humans and other intelligent animals:
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/super-turing-geom.html

---

Installed: 11 Mar 2018
Updated: 20 Jan 2019; 7 Feb 2019; 6 Apr 2024
27 May 2018; 19 Aug 2018; 2 Oct 2018; 24 Oct 2018;
More papers still to be included, annotated, etc.
This paper is
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/emotions-affect.html
This is part of the Birmingham Cognition and Affect (COGAFF) project:
http://www.cs.bham.ac.uk/research/projects/cogaff/

MORE RECENT ITEMS
(Main contents below)

Some of what follows is now explicitly or implicitly subsumed (and in some cases updated) in the more recent document mentioned above
https://www.cs.bham.ac.uk/research/projects/cogaff/misc/metamorphosis.html
(Still being updated.)

Work on evolution of consciousness, from its very simple (or precursor) life/proto-life forms (4.5 page abstract for invited talk at APA conference 14th Jan 2021):
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/sloman-apa-2021.pdf
The meta-configured genome project:
http://www.cs.bham.ac.uk/research/projects/cogaff/movies/meta-config/
Some of the ideas are updated in the metamorphosis.html document cited above.

Work in 2021 on the deep problem of explaining how chemical processes inside eggs can produce cognitively competent hatchlings (e.g. baby avocets), apparently with significant knowledge about where food is to be found and how to obtain it:
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/sloman-morcom.html

Work by Luc Beaudoin PhD Thesis 1994
Recent:
Learning from works of Art
https://cogzest.com/:
News from Luc Beaudoin's Cogzest project, continuing the work he began here in 1991 and

others listed below. Some of his more recent work https://cogzest.com/projects/.

On Monday 28 January 2019, two researchers in the School of Philosophy gave a talk in the School of Computer Science, on
Robots, Emotions, and Epistemic Rational Assessability
Slides for the talk: PDF with linked videos Rationality_Robot_emotions.pdf
Speakers: Matilde Aliffi and Helen Ryland
https://vpp.midlands3cities.ac.uk/display/mxa654bhamacuk/
https://www.birmingham.ac.uk/schools/ptr/departments/philosophy/research/postgraduateresearch/profiles/ryland-helen.aspx
(Both doctoral Researchers, Department of Philosophy, University of Birmingham)
Abstract:
There is a current lack of philosophical research on whether robots could have emotions. In this talk, we argue that the idea that a robot could have emotions is more plausible than currently assumed. We will demonstrate this by giving examples of robots that appear to have some of the emotional components that are usually involved in human emotional experiences. This opens up new philosophical questions specifically about the rational status of these robots? emotions. We claim that if a robot can have emotions, or ?robot-like emotions?, then these emotions may be open to epistemic rational assessment.

MAIN CONTENTS
Papers and notes on the Cognition and Affect Project

Related work done since mid 1960s, including later work with students and colleagues.
This work was first influenced by AI during 1972-3, when I spent a year in Edinburgh, having my brain rewired, for a new approach to philosophy.
(This list is still incomplete.)

External summaries
Here's a very short summary, with diagrams, of the H-Cogaff architecture schema, produced by someone I've never met
http://www.garfixia.nl/h-cogaff

INCOMPLETE CONTENTS LIST
(To be expanded)

Early papers relevant to emotions/motivation/affect/preferences/values...
1969
How to derive "better" from "is"
Sloman
1970
"Ought" and "Better"
Sloman
1978
1978 Book: The Computer Revolution in Philosophy

Papers more centrally concerned with varieties of affect
1981

You don't need a soft skin to have a warm heart
Sloman and Croucher
1981
Why robots will have emotions
Sloman and Croucher
1982
Towards a Grammar of Emotions
Sloman
1987
Motives Mechanisms and Emotions
Sloman
1990
Prolegomena to a Theory of Communication and Affect
Sloman
1991
A Proposal for a Study of Motive Processing
Luc Beaudoin (Thesis proposal)
1992
Appendix to JCI proposal, The Attention and Affect Project
Aaron Sloman and Glyn Humphreys
This paper was mostly written by the first author, although it is partly based on, and develops, ideas of the second author.
1992
What are the phenomena to be explained?
Sloman
1992 Towards an information processing theory of emotions
Sloman
1992
Silicon Souls, How to design a functioning mind
Sloman
(Professorial Inaugural Lecture, University of Birmingham 1992)
1993
The mind as a control system
Sloman
1993
A study of motive processing and attention,
Beaudoin and Sloman (April 1993)
1993
The Terminological Pitfalls of Studying Emotion
Tim Read
1994
Computational Modelling Of Motive-Management Processes
Sloman, Beaudoin, Wright ISRE 1994 Poster
1994
Goal processing in autonomous agents
Luc Beaudoin (PhD thesis)
1994
An Emotional Agent -- Detection and Control of Emergent States in an Autonomous

Resource-Bounded Agent
Ian Wright (Thesis proposal)
1995
Information about the SimAgent toolkit
Aaron Sloman and Riccardo Poli (later Brian Logan)
1995
Playing God: A toolkit for building agents
Information about the SimAgent toolkit
Aaron Sloman and Riccardo Poli
Date: November 1994 to March 1995
1995
SIM_AGENT: A toolkit for exploring agent designs
Aaron Sloman and Riccardo Poli
1996
Towards a Design-Based Analysis of Emotional Episodes,
(Grief paper.)
Ian P. Wright, Aaron Sloman, Luc P. Beaudoin,
1998
Cognition and affect: Architectures and tools
Brian Logan and Aaron Sloman
1998
Architectures and Tools for Human-Like Agents
Aaron Sloman and Brian Logan
1999
PhD Thesis Proposal: Distributed Reflective Architectures,
Catriona M. Kennedy
1999
Patrice Terrier interviews Aaron Sloman for EACE QUARTERLY
(August 1999)
1999
Title: Architectural Requirements for Human-like Agents Both Natural and Artificial.
(What sorts of machines can love?)
Aaron Sloman. Invited conference talk, later published in *Human Cognition And Social Agent Technology*
Ed. Kerstin Dautenhahn,
1999
How many separately evolved emotional beasties live within us?
Aaron Sloman
Invited Talk: at workshop on *Emotions in Humans and Artifacts* Vienna, August 1999
Final version published 2002.
1999-2000
Evolvable architectures for human-like minds
Aaron Sloman and Brian Logan
Invited talk at 13th Toyota Conference, on "Affective Minds" Nagoya Japan, Nov-Dec 1999
Published in *Affective Minds,* Ed. Giyoo Hatano, Elsevier, October 2000
2003
Progress report on the Cognition and Affect project:
Architectures, Architecture-Schemas, And The New Science of Mind

Aaron Sloman
2004
AAAI 2004 Workshop invited talk: What are emotion theories about?
Aaron Sloman
2004
Simulating Infant-Carer Relationship Dynamics
Dean Petters
2004
How to Determine the Utility of Emotions (At AAAI-04)
Matthias Scheutz
2005
The Architectural Basis of Affective States and Processes
Sloman, Chrisley and Scheutz: invited book chapter for "Who needs emotions" (eds. Arbib and Fellous).
2006 2009
Architecture-Based Motivation vs Reward-Based Motivation
Strongly challenges almost all published theories of motivation, especially in experimental psychology, neuroscience, and AI, and some in philosophy. They all grossly over-simplify the biological facts.
Aaron Sloman
2017
Architectures underlying cognition and affect in natural and artificial systems
(Extended Abstract for invited talk at AISB 2017)
2017
Cognition and Affect: Past and Future
Cognition and Affect Workshop, Following AISB 2017 Discussions.
University of Birmingham, 24th April 2017
2017
Anger, an example of conceptual analysis,
(Background material for workshop)
Aaron Sloman

MORE TO BE ADDED

REFERENCES AND LINKS

BACK TO CONTENTS

Early papers relevant to emotions/motivation/affect/preferences/values...

- **1969**
- http://www.cs.bham.ac.uk/research/cogaff/sloman-better.html
  http://www.cs.bham.ac.uk/research/cogaff/sloman-better.pdf
  Title: How to derive "better" from "is"
  Author: Aaron Sloman
  Originally Published as: How to derive "better" from "is"

American Philosophical Quarterly, Vol 6, Number 1, Jan 1969, pp 43-52.
This aims to show how subjective evaluations expressed using "better" and related words, may originally have been rooted in factual comparisons of alternative ways of meeting a need or serving a goal.
Online here: http://www.cs.bham.ac.uk/research/projects/cogaff/62-80.html#1969-02
Abstract:

ONE type of naturalistic analysis of words like "good," "ought," and "better" defines them in terms of criteria for applicability which vary from one context to another (as in "good men," "good typewriter," "good method of proof"), so that their meanings vary with context. Dissatisfaction with this "crude" naturalism leads some philosophers to suggest that the words have a context-independent non-descriptive meaning defined in terms of such things as expressing emotions, commanding, persuading, or guiding actions.

There are well-known objections to both approaches, and the aim of this paper is to suggest an alternative which has apparently never previously been considered, for the very good reason that at first sight it looks so unpromising, namely the alternative of defining the problematic words as logical constants.

This should not be confused with the programme of treating them as undefined symbols in a formal system, which is not new. In this essay an attempt will be made to define a logical constant "Better" which has surprisingly many of the features of the ordinary word "better" in a large number of contexts. It can then be shown that other important uses of "better" may be thought of as derived from this use of the word as a logical constant.

The new symbol is a logical constant in that its definition (i.e., the specification of formation rules and truth-conditions for statements using it) makes use only of such concepts as "entailment," "satisfying a condition," "relation," "set of properties," which would generally be regarded as purely logical concepts. In particular, the definition makes no reference to wants, desires, purposes, interests, prescriptions, choice, non-descriptive uses of language, and the other paraphernalia of non-naturalistic (and some naturalistic) analyses of evaluative words.

(However, some of those 'paraphernalia' can be included in arguments/subjects to which the complex relational predicate 'better' is applied.)

- **1970**
http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-ought-and-better.html
http://www.cs.bham.ac.uk/research/projects/cogaff/ought-better.pdf
Title: 'Ought' and 'Better'
Author: Aaron Sloman
Originally published as: A. Sloman, 'Ought and Better'
Mind, vol LXXIX, No 315, July 1970, pp 385--394)
Abstract:

This is a sequel to the 1969 paper on "How to derive 'Better' from 'Is'". It presupposes the analysis of 'better' in the earlier paper, and argues that statements using the word 'ought' say something about which of a collection of alternatives is better than the others, in contrast with statements using 'must' or referring to 'obligations', or what is 'obligatory'.

> The underlying commonality between superficially different statements like 'You should take an umbrella with you' and 'The sun should come out soon' is explained, along with some other philosophical puzzles, e.g. concerning why 'ought' does not imply 'can', contrary to what some philosophers have claimed.

1978

- **1978 Book: The Computer Revolution in Philosophy: Philosophy Science and models of mind**
  This is now freely available online. It includes sections relevant to emotions and other varieties of affect, including a chapter on conceptual analysis, a chapter on architectures and final chapter related to consciousness, including remarks on emotions. The revised online edition (2001--onwards) (html or pdf) is searchable. E.g. search for occurrences of "attitude", "desire", "emotion", "emotional", "feeling", "interrupt", "mood", "motive", "preference", "surprise", "architecture", etc.
  http://www.cs.bham.ac.uk/research/projects/cogaff/crp/

Papers more centrally concerned with varieties of affect

1981

- **Title: You don't need a soft skin to have a warm heart:**
  Towards a computational analysis of motives and emotions
  Aaron Sloman and Monica Croucher
  September 1981 University of Sussex
  sloman-croucher-warm-heart.html
  sloman-croucher-warm-heart.pdf
  Authors: Aaron Sloman and Monica Croucher

  > Originally a *Cognitive Science Research Paper* at Sussex University:
  > Sloman, Aaron and Monica Croucher, "You don't need a soft skin to have a warm heart: towards a computational analysis of motives and emotions," CSRP 004, 1981.
  > (Written circa 1980-81, at Sussex University: CSRP 004, 1981.)
  >
  > This was submitted unsuccessfully to *Behavioural and Brain Sciences Journal*.

  Abstract:
  The paper introduces an interdisciplinary methodology for the study of minds of animals humans and machines, and, by examining some of the pre-requisites for intelligent decision-making, attempts to provide a framework for integrating some of the fragmentary studies to be found in Artificial Intelligence.

  The space of possible architectures for intelligent systems is very large. This essay takes steps towards a survey of the space, by examining some environmental and functional constraints, and discussing mechanisms capable of fulfilling them. In particular, we examine a subspace close to the human mind, by illustrating the variety of motives to be expected in a human-like system, and types of processes they can produce in meeting some of the constraints.

This provides a framework for analysing emotions as computational states and processes, and helps to undermine the view that emotions require a special mechanism distinct from cognitive mechanisms. The occurrence of emotions is to be expected in any intelligent robot or organism able to cope with multiple motives in a complex and unpredictable environment.

Analysis of familiar emotion concepts (e.g. anger, embarrassment, elation, disgust, pity, etc.) shows that they involve interactions between motives (e.g. wants, dislikes, ambitions, preferences, ideals, etc.) and beliefs (e.g. beliefs about the fulfilment or violation of a motive), which cause processes produced by other motives (e.g. reasoning, planning, execution) to be disturbed, disrupted or modified in various ways (some of them fruitful). This tendency to disturb or modify other activities seems to be characteristic of all emotions. In order fully to understand the nature of emotions, therefore, we need to understand motives and the types of processes they can produce.

This in turn requires us to understand the global computational architecture of a mind. There are several levels of discussion: description of methodology, the beginning of a survey of possible mental architectures, speculations about the architecture of the human mind, analysis of some emotions as products of the architecture, and some implications for philosophy, education and psychotherapy.

- Aaron.Sloman_why_robot_emotions.pdf
  Title: Why robots will have emotions
  Authors: Aaron Sloman and Monica Croucher
  Date: August 1981
  Originally appeared in **Proceedings IJCAI 1981**, Vancouver
  Also Sussex University Cognitive Science Research paper No 176
  Abstract:

  > Emotions involve complex processes produced by interactions between motives, beliefs, percepts, etc. E.g. real or imagined fulfilment or violation of a motive, or triggering of a ’motive-generator’, can disturb processes produced by other motives. To understand emotions, therefore, we need to understand motives and the types of processes they can produce. This leads to a study of the global architecture of a mind. Some constraints on the evolution of minds are discussed. Types of motives and the processes they generate are sketched.
  >
  > (Note we now use slightly different terminology from that used in this paper. In particular, what the paper labelled as "intensity" we now call "insistence", i.e. the capacity to divert attention from other things.)
  >
  > **NB**
  > This paper is often misquoted as arguing that robots (or at least intelligent robots) *should* have emotions. On the contrary, the paper argues that certain sorts of high level disturbances (i.e. emotional states) will be capable of arising out of interactions between mechanisms that exist for other reasons. Similarly ’thrashing’ is capable of occurring in multi-processing operating systems that support swapping and paging, but that does not mean that operating systems *should* produce thrashing.

A more recent analysis of the confused but fashionable arguments (e.g. based on Damasio's writings) claiming that emotions are needed for intelligence can be found in this semi-popular presentation.

One of the arguments is analogous to arguing that a car requires a functioning horn for its starter motor to work, because damaging the battery can disable the horn and disable the starter motor.

1982

- **Sloman.emot.gram.pdf**
  Title: Towards a Grammar of Emotions,
  Invited paper in **New Universities Quarterly**, 36,3, pp 230-238, 1982.
  Authors: Aaron Sloman
  Date: Installed here 6 Dec 1998 (Originally Published in 1982)

  Abstract:
  By analysing what we mean by 'A longs for B', and similar descriptions of emotional states we see that they involve rich cognitive structures and processes, i.e. computations. Anything which could long for its mother, would have to have some sort of representation of its mother, would have to believe that she is not in the vicinity, would have to be able to represent the possibility of being close to her, would have to desire that possibility, and would have to be to some extent pre-occupied or obsessed with that desire. The paper includes a fairly detailed discussion of what it means to say 'X is angry with Y', and relationships between anger, exasperation, annoyance, dismay, etc. Emotions are contrasted with attitudes and moods.
  NOTE:
  This paper contains examples of the technique of conceptual analysis explained in a tutorial that formed Chapter 4 of *The Computer Revolution in Philosophy* (1978)
  That chapter is available as part of the new online edition of the book:
  http://www.cs.bham.ac.uk/research/projects/cogaff/crp/#chap4

1987

- **Aaron.Sloman_Motives.Mechanisms.pdf**
  Aaron.Sloman_Motives.Mechanisms.txt
  Title: Motives Mechanisms and Emotions
  Author: Aaron Sloman
  Date: 1987
  In **Cognition and Emotion** 1,3, pp.217-234 1987, later reprinted in M.A. Boden (ed) **The Philosophy of Artificial Intelligence**, "Oxford Readings in Philosophy" Series Oxford University Press, pp 231-247 1990.
  (Previously available as Cognitive Science Research Paper No 62, Sussex University.)
  Extract from introduction

  > Ordinary language makes rich and subtle distinctions between different sorts of mental states and processes such as mood, emotion, attitude, motive, character, personality, and so on. Our words and concepts have been honed for centuries against the intricacies of

real life under pressure of real needs and therefore give deep hints about the human mind. Yet actual usage is inconsistent, and our ability to articulate the distinctions we grasp and use intuitively is as limited as our ability to recite rules of English syntax. Words like "motive" and "emotion" are used in ambiguous and inconsistent ways. The same person will tell you that love is an emotion, that she loves her children deeply, and that she is not in an emotional state. Many inconsistencies can be explained away if we rephrase the claims using carefully defined terms. As scientists we need to extend colloquial language with theoretically grounded terminology that can be used to mark distinctions and describe possibilities not normally discerned by the populace. For instance, we'll see that love is an attitude, not an emotion, though deep love can easily trigger emotional states. In the jargon of philosophers (Ryle 1949), attitudes are dispositions, emotions are episodes, though with dispositional elements. For a full account of these episodes and dispositions we require a theory about how mental states are generated and controlled and how they lead to action -- a theory about the mechanisms of mind. The theory should explain how internal representations are built up, stored, compared, and used to make inferences, formulate plans or control actions. Outlines of a theory are given. Design constraints for intelligent animals or machines are sketched, then design solutions are related to the structure of human motivation and to computational mechanisms underlying familiar emotional states.

1990

- **sloman-prolegomena-communication-affect.pdf (PDF)**
  sloman-prolegomena-communication-affect.html (HTML)
  Author: Aaron Sloman
  Title: Prolegomena to a Theory of Communication and Affect
  In Ortony, A., Slack, J., and Stock, O. (Eds.) **Communication from an Artificial Intelligence Perspective: Theoretical and Applied Issues.** Heidelberg, Germany: Springer, 1992, pp 229-260.
  (HTML version added 23 May 2015)

  Invited paper presented, Nov 1990, to NATO Advanced Research Workshop on "Computational theories of communication and their applications: Problems and Prospects". Originally available as Cognitive Science Research Paper, CSRP-91-05, School of Computer Science, University of Birmingham.

  Abstract:
  As a step towards comprehensive computer models of communication, and effective human machine dialogue, some of the relationships between communication and affect are explored. An outline theory is presented of the architecture that makes various kinds of affective states possible, or even inevitable, in intelligent agents, along with some of the implications of this theory for various communicative processes. The model implies that human beings typically have many different, hierarchically organised, dispositions capable of interacting with new information to produce affective states, distract attention, interrupt ongoing actions, and so on. High "insistence" of motives is defined in relation to a tendency to penetrate an attention filter mechanism, which seems to account for the partial loss of control involved in emotions. One conclusion is that emulating human communicative abilities will not be achieved easily. Another is that it will be even more difficult to design and build computing systems that reliably achieve interesting communicative goals.

1991

- BeaudoinSloman-1991-proposalForStudyOfMotiveProcessing.pdf
Title: A Proposal for a Study of Motive Processing
Authors: Luc Beaudoin and Aaron Sloman
Date: 1991
Date Installed here: 30 Jan 2016

  Where published: PhD Thesis proposal Luc Beaudoin, University of Birmingham

  Abstract:

  This paper was mostly written by the first author, although it is based on and develops ideas of the second author. The nursemaid scenario was first described by the second author (Sloman, 1986). At the time of writing Luc Beaudoin was in the process of implementing the model described in the paper.

  In this paper we discuss some of the essential features and context of human motive processing, and we characterize some of the state transitions of motives. We then describe in detail a domain for designing an agent exhibiting some of these features. Recent related work is briefly reviewed to demonstrate the need for extending theories to account for the complexities of motive processing described here.

  The nursemaid scenario is available at
  http://www.cs.bham.ac.uk/research/projects/cogaff/misc/nursemaid-scenario.html

  Luc's work growing out of this proposal, and his thesis, are referenced below. Search for "Beaudoin".

1992

- sloman-humphreys-jci-proposal.pdf
Title: Appendix to JCI proposal, The Attention and Affect Project
Authors: Aaron Sloman and Glyn Humphreys
Date: January 1992
Appendix to research grant proposal for the Attention and Affect project. (Paid for computer and computer officer support, and some workshops, for three years, funded by UK Joint Research Council initiative in Cognitive Science and HCI, 1992-1995.)
Later this grew into the Birmingham Cognition and Affect (CogAff) project.

- Aaron.Sloman_Phenomena.Explain.pdf
Title: What are the phenomena to be explained?
Author: Aaron Sloman
Date: Dec 1992

  Seminar notes for the *Attention and Affect* Project, later re-named *Cognition and Affect*, summarising its long term objectives better.

- **Aaron.Sloman_IP.Emotion.Theory.pdf**
  Title: Towards an information processing theory of emotions
  Author: Aaron Sloman
  Date: Dec 1992
  More seminar notes for the Attention and Affect Project

- Aaron.Sloman_Silicon.Souls.pdf
  Title: Silicon Souls, How to design a functioning mind (Inaugural lecture)
  Author: Aaron Sloman
  Date: May 1992
  Professorial Inaugural Lecture, Birmingham, May 1992 In the form of lecture slides for an excessively long lecture. Much of this is replicated and expanded in other papers published since.

1993

- Aaron.Sloman_Mind.as.controlsystem/ (HTML)
  New PDF derived from new HTML:
  Aaron.Sloman_Mind.as.controlsystem.pdf
  Title: The Mind as a Control System,
  Author: Aaron Sloman
  In *Philosophy and the Cognitive Sciences*, (eds) C. Hookway and D. Peterson, Cambridge University Press, pp 69--110
  Date: 1993
  Originally Presented at Royal Institute of Philosophy conference on Philosophy and the Cognitive Sciences, in Birmingham in 1992, with proceedings published later.
  Abstract:

  > Many people who favour the design-based approach to the study of mind, including the author previously, have thought of the mind as a computational system, though they don’t all agree regarding the forms of computation required for mentality. Because of ambiguities in the notion of ’computation’ and also because it tends to be too closely linked to the concept of an algorithm, it is suggested in this paper that we should rather construe the mind (or an agent with a mind) as a control system involving many interacting control loops of various kinds, most of them implemented in high level virtual machines, and many of them hierarchically organised. (Some of the sub-processes are clearly computational in character, though not necessarily all.) A feature of the system is that the same sensors and motors are shared between many different functions, and sometimes they are shared concurrently, sometimes sequentially. A number of implications are drawn out, including the implication that there are many informational substates, some incorporating factual information, some control information, using diverse forms of representation. The notion of architecture, i.e. functional differentiation into interacting components, is explained, and the conjecture put forward that in order to account for the main characteristics of the human mind it is more important to get the architecture right than to get the mechanisms right (e.g. symbolic vs neural mechanisms). Architecture dominates mechanism.
  > (In 2018 I began to revise this opinion because of problems of explaining deep ancient mathematical competences by means of computer models: here.)

1993

- **Aaron.Sloman_prospects.pdf**
  Title: Prospects for AI as the General Science of Intelligence
  Author: Aaron Sloman
  Date: April 1993

  in *Proceedings AISB93*, published by IOS Press as a book: *Prospects for Artificial Intelligence*
  Editors: A.Sloman, D.Hogg, G.Humphreys, D. Partridge, A. Ramsay

  Abstract:
  Three approaches to the study of mind are distinguished: semantics-based, phenomena-based and design-based. Requirements for the design-based approach are outlined. It is argued that AI as the design-based approach to the study of mind has a long future, and pronouncements regarding its failure are premature, to say the least.

1993

- **Luc.Beaudoin.and.Sloman_Motive_proc.pdf**
  Title: A study of motive processing and attention,
  Authors: Luc P. Beaudoin and Aaron Sloman
  in *Proceedings AISB93*, published by IOS Press as a book:
  Prospects for Artificial Intelligence
  A.Sloman, D.Hogg, G.Humphreys, D. Partridge, A. Ramsay (eds)

Date: April 1993
Abstract:

> We outline a design based theory of motive processing and attention, including: multiple motivators operating asynchronously, with limited knowledge, processing abilities and time to respond. Attentional mechanisms address these limits using processes differing in complexity and resource requirements, in order to select which motivators to attend to, how to attend to them, how to achieve those adopted for action and when to do so. A prototype model is under development. Mechanisms include: motivator generators, attention filters, a dispatcher that allocates attention, and a manager. Mechanisms like these might explain the partial loss of control of attention characteristic of many emotional states.

1993

- Tim.Read-et.al_TerminlogyPit.pdf
  Title: The Terminological Pitfalls of Studying Emotion
  Authors: Tim Read (Research seminar paper)
  Date: Aug 1993

  Abstract:
  The research community is full of papers with titles that include terms like ’emotion’, ’motivation’, ’cognition’, and ’attention’. However when these terms are used they are either considered to be so obvious as not to warrant a definition, or are defined in overly simplistic and arbitrary ways. The reasons behind our usage of existing terminology is easy to see, but the problems inherent with it are not. The use of such terminology gives rise to a whole set of problems, chief among them are confusion and pointless semantic disagreement. These problems occur because the current terminology is too vague, and burdened with acquired meaning. We need to replace it with terminology that emerges from a putatively complete theory of the conceptual space of mechanisms and behaviours, spanning several functional levels (e.g.: neural, behavioural and computational). Research that attempts to use the current terminology to build larger and more complex theory, just adds to the existing confusion. In this paper I examine the reasons behind the use of current terminology, explore the problems inherent with it, and offer a way to resolve these problems. The days when one small research team could hope to produce a theory to explain the complete range of phenomena currently referred to as being ’emotional’ have passed. It is time for concerted and coordinated activity to understand the relation of mechanisms to behaviour. This will give rise to clear and unambiguous terminology that is defined at different functional levels. Until the current terminological problems are solved, our rate of progress will be slow.

1994

- Aaron.Sloman_isre.pdf
  Title: Computational Modelling Of Motive-Management Processes
  "Poster" prepared for the Conference of the International Society for Research in Emotions, Cambridge July 1994.
  Authors: Aaron Sloman, Luc Beaudoin and Ian Wright
  Revised version in *Proceedings ISRE94*, edited by Nico Frijda, ISRE Publications.

Date: 29 July 1994 (PDF version added here 25 Dec 2005)

Abstract:
This is a 5 page summary with three diagrams of the main objectives and some work in progress at the University of Birmingham Cognition and Affect project. involving: Professor Glyn Humphreys (School of Psychology), and Luc Beaudoin, Chris Paterson, Tim Read, Edmund Shing, Ian Wright, Ahmed El-Shafei, and (from October 1994) Chris Complin (research students). The project is concerned with "global" design requirements for coping simultaneously with coexisting but possibly unrelated goals, desires, preferences, intentions, and other kinds of motivators, all at different stages of processing. Our work builds on and extends seminal ideas of H.A.Simon (1967). We are exploring "broad and shallow" architectures combining varied capabilities most of which are not implemented in great depth. The poster summarises some ideas about management and meta-management processes, attention filtering, and the relevance to emotional states involved "perturbances", where there is partial loss of control of attention.

1994-5

- http://www.cs.bham.ac.uk/research/projects/cogaff/Luc.Beaudoin_thesis.pdf
Title: Goal processing in autonomous agents (PhD thesis)
Author: Luc P. Beaudoin
Date: 31 Aug 1994 (Updated March 13th 1995)
(PDF version added 18 May 2003; PDF Corrected 24 Dec 2014)
A thesis submitted to the Faculty of Science of the University of Birmingham for the degree of PhD in Cognitive Science. (Supervisor: Aaron Sloman).
Abstract:
The objective of this thesis is to elucidate goal processing in autonomous agents from a design-stance. A. Sloman's theory of autonomous agents is taken as a starting point (Sloman, 1987; Sloman, 1992b). An autonomous agent is one that is capable of using its limited resources to generate and manage its own sources of motivation. A wide array of relevant psychological and AI theories are reviewed, including theories of motivation, emotion, attention, and planning. A technical yet rich concept of goals as control states is expounded. Processes operating on goals are presented, including vigilational processes and management processes. Reasons for limitations on management parallelism are discussed. A broad design of an autonomous agent that is based on M. Georgeff's (1986) Procedural Reasoning System is presented. The agent is meant to operate in a microworld scenario. The strengths and weaknesses of both the design and the theory behind it are discussed. The thesis concludes with suggestions for studying both emotion ("perturbance") and pathologies of attention as consequences of autonomous goal processing.

1994

- Ian.Wright_emotional_agent.pdf
Title: An Emotional Agent --
The Detection and Control of Emergent States in an Autonomous Resource-Bounded Agent
(PhD Thesis Proposal)
Date: October 31 1994
Author: Ian Wright
Abstract:
In dynamic and unpredictable domains, such as the real world, agents are continually faced

with new requirements and constraints on the quality and types of solutions they produce. Any agent design will always be limited in some way. Such considerations highlight the need for self-referential mechanisms, i.e. agents with the ability to examine and reason about their internal processes in order to improve and control their own functioning.

This work aims to implement a prototype agent architecture that meets the requirements for self-referential systems, and is able to exhibit perturbant ('emotional') states, detect such states and attempt to do something about them. Results from this research will contribute to autonomous agent design, emotionality, internal perception and meta-level control; in particular, it is hoped that we will
i. provide a (partial) implementation of Sloman's theory of perturbances (Sloman, 81) within the NML1 design (Beaudoin, 94),
ii. investigate the requirements for the self-detection and control of processing states, and
iii. demonstrate the adaptiveness of, the need for, and consequences of, self-control mechanisms that meet the requirements for self-referential systems.

1995

- **http://www.cs.bham.ac.uk/research/projects/cogaff/sim_agent.pdf**
  Title: Playing God: A toolkit for building agents
  Information about the SimAgent toolkit
  Authors: Aaron Sloman and Riccardo Poli
  Date: November 1994 to March 1995
  Part of the early online documentation for the SimAgent toolkit.
  Later extended with Brian Logan, to 2001
  More recent Package documentation:
  http://www.cs.bham.ac.uk/research/projects/poplog/packages/simagent.html
  Link to the main SimAgent overview page. Includes some (ancient) movies demonstrating simple uses of the toolkit.
  Code and documentation online:
  http://www.cs.bham.ac.uk/research/projects/poplog/sim
  Abstract:
  These files give partial descriptions of the SimAgent toolkit, implemented in Poplog Pop-11, based on the Poprulebase package, providing multiple, concurrent, interacting, rule-based systems, for exploring architectures for individual or interacting agents.
  See also the Atal95 paper
  Aaron.Sloman_Riccardo.Poli_sim_agent_toolkit.pdf

1995

- **http://www.cs.bham.ac.uk/research/projects/cogaff/Aaron.Sloman_Riccardo.Poli_sim_agent_toolkit.pdf**
  Title: SIM_AGENT: A toolkit for exploring agent designs
  Authors: Aaron Sloman and Riccardo Poli
  In **Intelligent Agents Vol II (ATAL-95)**,
  Eds. Mike Wooldridge, Joerg Mueller, Milind Tambe, Springer-Verlag 1996 pp 392--407.
  Updated version of: Cognitive Science technical report: CSRP-95-3 School of Computer Science, the University of Birmingham.
  Presented at ATAL-95, Workshop on Agent Theories, Architectures, and Languages, at

IJCAI-95 Workshop, Montreal, August 1995
Date: Oct 1995
Abstract:
SIM_AGENT is a toolkit that arose out of a project concerned with designing an architecture for an autonomous agent with human-like capabilities. Analysis of requirements showed a need to combine a wide variety of richly interacting mechanisms, including independent asynchronous sources of motivation and the ability to reflect on which motives to adopt, when to achieve them, how to achieve them, and so on. These internal 'management' (and meta-management) processes involve a certain amount of parallelism, but resource limits imply the need for explicit control of attention. Such control problems can lead to emotional and other characteristically human affective states. In order to explore these ideas, we needed a toolkit to facilitate experiments with various architectures in various environments, including other agents. The paper outlines requirements and summarises the main design features of a Pop-11 toolkit supporting both rule-based and 'sub-symbolic' mechanisms. Some experiments including hybrid architectures and genetic algorithms are summarised. (More recent documentation on the toolkit is below.)

1996

- **Grief Paper**
  Title: Towards a Design-Based Analysis of Emotional Episodes,
  Authors: Ian P. Wright, Aaron Sloman, Luc P. Beaudoin,
  Philosophy Psychiatry and Psychology, 3, 2, pp. 101--126, 1996,
  http://www.cs.bham.ac.uk/research/projects/cogaff/96-99.html#2
  With several invited commentaries.
  This (invited) journal paper takes a real example of long-term grief as a case study that conflicts with most published theories/models of emotion, e.g. because grief can co-exist with many other states, "waiting in the wings, ready to pounce on the slightest provocation" e.g. seeing a reminder of the lost child.

  Date: Oct 1995 (published 1996)

  Appeared (with commentaries) in **Philosophy Psychiatry and Psychology**, vol 3 no 2, 1996, pp 101--126.

  Journal web site:
  http://muse.jhu.edu/journals/philosophy_psychiatry_and_psychology/v003/3.2wright01.html

  The commentaries, by
  - Dan Lloyd,
  - Cristiano Castelfranchi and Maria Miceli
  - Margaret Boden

  are available here, followed by a reply by the authors:
  http://muse.jhu.edu/journals/philosophy_psychiatry_and_psychology/toc/ppp3.2.html

  (This is a revised version of the paper presented to the Geneva Emotions Workshop, April 1995, entitled
  "The Architectural Basis for Grief".)

Abstract:

The design-based approach is a methodology for investigating mechanisms capable of generating mental phenomena, whether introspectively or externally observed, and whether they occur in humans, other animals or robots. The study of designs satisfying requirements for autonomous agency can provide new deep theoretical insights at the information processing level of description of mental mechanisms. Designs for working systems (whether on paper or implemented on computers) can systematically explicate old explanatory concepts and generate new concepts that allow new and richer interpretations of human phenomena. To illustrate this, some aspects of human grief are analysed in terms of a particular *information processing architecture* being explored in our research group.

We do not claim that **this** architecture is part of the causal structure of the human mind; rather, it represents an early stage in the iterative search for a deeper and more general architecture, capable of explaining more phenomena. However even the current early design provides an interpretative ground for some familiar phenomena, including characteristic features of certain emotional episodes, particularly the phenomenon of **perturbance** (a partial or total loss of control of attention).

The paper attempts to expound and illustrate the design-based approach to cognitive science and philosophy, to demonstrate the potential effectiveness of the approach in generating interpretative possibilities, and to provide first steps towards an information processing account of 'perturbant', emotional episodes.

1998

- **http://www.cs.bham.ac.uk/research/projects/cogaff/logan-sloman-aa98poster.pdf**
Title: Cognition and affect: Architectures and tools
Authors: Brian Logan and Aaron Sloman
Date: Feb 1998
Summary of poster presentation. In Proceedings of the Second International Conference on Autonomous Agents (Agents '98), ACM Press, 1998, pp 471--472.

Abstract:
Which agent architectures are capable of justifying descriptions in terms of the 'higher level' mental concepts applicable to human beings? We propose a new kind of architecture-based semantics for mentalistic descriptions in which mental concepts (e.g. 'believes', 'desires', 'intends', 'mood', 'emotion', etc.) are grounded in assumptions about information processing architectures, and not merely in concepts based solely on Dennett's 'intentional stance'. These ideas have led to the design of the SIM_AGENT toolkit which has been used to explore a variety of such architectures.

1998

- **http://www.cs.bham.ac.uk/research/projects/cogaff/Sloman.and.Logan.eccm98.pdf**
Title: Architectures and Tools for Human-Like Agents
Authors: Aaron Sloman and Brian Logan
Date: 11 Mar 1998

In *Proceedings 2nd European Conference on Cognitive Modelling*, Nottingham, April 1-4, 1998. Eds Frank Ritter and Richard M. Young, Nottingham University Press, pp 58--65. 1998
Abstract:
This paper discusses agent architectures which are describable in terms of the "higher level" mental concepts applicable to human beings, e.g. "believes", "desires", "intends" and "feels". We conjecture that such concepts are grounded in a type of information processing architecture, and not simply in observable behaviour nor in Newell's knowledge-level concepts, nor Dennett's "intentional stance." A strategy for conceptual exploration of architectures in design-space and niche-space is outlined, including an analysis of design trade-offs. The SIM_AGENT (SimAgent) toolkit, developed to support such exploration, including hybrid architectures, is described briefly.

1999

- **http://www.cs.bham.ac.uk/research/projects/cogaff/Kennedy.proposal.pdf**
  Title: PhD Thesis Proposal: Distributed Reflective Architectures
  Author:: Catriona M. Kennedy
  Date: 23 July 1999

  Abstract:

  > The autonomy of a system can be defined as its capability to recover from unforeseen difficulties without any user intervention. This thesis proposal addresses a small part of this problem, namely the detection of anomalies within a system's own operation by the system itself. It is a response to a challenge presented by immune systems which can distinguish between "self " and "nonself ", i.e. they can recognise a "foreign" pattern (due to a virus or bacterium) as different from those associated with the organism itself, even if the pattern was not previously encountered. The aim is to apply this requirement to an artificial system, where "nonself " may be any form of deliberate intrusion or random anomalous behaviour due to a fault. When designing reflective architectures or self-diagnostic systems, it is simpler to rely on a single coordination mechanism to make the system work as intended. However, such a coordination mechanism cannot be inspected or repaired by the system itself, which means that there is a gap in its reflective coverage. To try to overcome this limitation, this thesis proposal suggests a conceptual frame-work based on a network of agents where each agent monitors the whole network from a unique and independent perspective and where the perspectives are not globally "managed". Each agent monitors the fault-detection capability and control algorithms of other agents (a process called meta-observation). In this way, the agents can collectively achieve reflective coverage of failures.

1999

- **http://www.cs.bham.ac.uk/research/projects/cogaff/Sloman.eace-interview.html**
  Title: Patrice Terrier interviews Aaron Sloman for EACE QUARTERLY
  Date: 3 Sep 1999

  Abstract:
  Patrice Terrier asks and Aaron Sloman attempts to answer questions about AI, about emotions, about the relevance of philosophy to AI, about Poplog, Sim_agent and other tools.

(EACE = European Association for Cognitive Ergonomics)

1999

- **http://www.cs.bham.ac.uk/research/projects/cogaff/sloman.vienna99.pdf**
  Title: How many separately evolved emotional beasties live within us?
  Revised version of Invited Talk: at workshop on *Emotions in Humans and Artifacts* Vienna, August 1999
  Final version published in *Emotions in Humans and Artifacts*, Eds Robert Trappl, Paolo Petta, and Sabine Payr, MIT Press, Cambridge MA., 2002
  Author: Aaron Sloman
  Date: 27 May 2000 (Revised: 8 Sep 2006}

  > The version installed here on 8th September 2006 has a few minor changes, including using the word 'CogAff' as a label for an *architecture schema* not an *architecture*, using the label 'H-cogaff' for the special case of the proposed human-like architecture, using 'ecosystem' instead of 'ecology', and an improved version of figure 11.

  Abstract:
  A problem which bedevils the study of emotions, and the study of consciousness, is that we assume a shared understanding of many everyday concepts, such as 'emotion', 'feeling', 'pleasure', 'pain', 'desire', 'awareness', etc. Unfortunately, these concepts are inherently very complex, ill-defined, and used with different meanings by different people. Moreover this goes unnoticed, so that people think they understand what they are referring to even when their understanding is very unclear. Consequently there is much discussion that is inherently vague, often at cross-purposes, and with apparent disagreements that arise out of people unwittingly talking about different things. We need a framework which explains how there can be all the diverse phenomena that different people refer to when they talk about emotions and other affective states and processes. The conjecture on which this paper is based is that adult humans have a type of information-processing architecture, with components which evolved at different times, including a rich and varied collection of components whose interactions can generate all the sorts of phenomena that different researchers have labelled "emotions". Within this framework we can provide rational reconstructions of many everyday concepts of mind. We can also allow a variety of different architectures, found in children, brain damaged adults, other animals, robots, software agents, etc., where different architectures support different classes of states and processes, and therefore different mental ontologies. Thus concepts like 'emotion', 'awareness', etc. will need to be interpreted differently when referring to different architectures. We need to limit the class of architectures under consideration, since for any class of behaviours there are indefinitely many architectures which can produce those behaviours. One important constraint is to consider architectures which might have been produced by biological evolution. This leads to the notion of a human architecture composed of many components which evolved under the influence of the other components as well as environmental needs and pressures. From this viewpoint, a mind is a kind of {\em ecosystem} (previously described as an 'ecology') of co-evolved sub-organisms acquiring and using different kinds of information and processing it in different ways, sometimes cooperating with one another and sometimes competing. Within this framework we can hope to study not only mechanisms underlying affective states and processes, but also other mechanisms which are often studied in isolation, e.g. vision, action mechanisms, learning mechanisms, 'alarm' mechanisms, etc. We can also explain why some models, and corresponding conceptions of

emotion, are shallow whereas others are deeper. Shallow models may be of practical use, e.g. in entertainment and interface design. Deeper models are required if we are to understand what we are, how we can go wrong, etc. This paper is a snapshot of a long term project addressing all these issues.

1999
http://www.cs.bham.ac.uk/research/projects/cogaff/Sloman.Logan.cacm.pdf
Title: Building cognitively rich agents using the SIM_AGENT toolkit,
(Invited paper in **Communications of the Association of Computing Machinery**, March 1999, vol 43, no 2, pp. 71-77;
Online (with inset report written by two users of the toolkit) at
http://portal.acm.org/citation.cfm?id=295704
and also (with inserted papers by uses):
http://www.cs.bham.ac.uk/research/projects/cogaff/Sloman.Logan-cacm-orig.pdf
Authors: Aaron Sloman and Brian Logan
Date: 17 Jan 1999

Abstract:

> An overview of some of the motivation of our research and design criteria for the SIM_AGENT toolkit for a special issue of CACM on multi-agent systems, edited by Anupam Joshi and Munindar Singh. Includes an insert by users: Jeremy Baxter and Richard Hepplewhite.
>
> For more information about the toolkit (now referred to as SimAgent), including movies of demos, see
> http://www.cs.bham.ac.uk/research/projects/poplog/packages/simagent.html

---

1999-2000

- **http://www.cs.bham.ac.uk/research/projects/cogaff/Sloman.kd.pdf**
Title: Architectural Requirements for Human-like Agents Both Natural and Artificial. (What sorts of machines can love?)
Later published in **Human Cognition And Social Agent Technology** Ed. Kerstin Dautenhahn, in the "Advances in Consciousness Research" series, John Benjamins Publishing
Extended version of slides on love for "Voice box" talk, presented in London
Authors: Aaron Sloman
Date: 10 Jan 1999 (Book Published, March 2000)

Abstract:
This paper, an expanded version of a talk on love given to a literary society, attempts to analyse some of the architectural requirements for an agent which is capable of having primary, secondary and tertiary emotions, including being infatuated or in love. It elaborates on work done previously in the Birmingham Cognition and Affect group, describing our proposed three level architecture (with reactive, deliberative and meta-management layers), showing how different sorts of emotions relate to those layers.

Some of the relationships between emotional states involving partial loss of control of attention (e.g. emotional states involved in being in love) and other states which involve dispositions (e.g. attitudes such as loving) are discussed and related to the architecture.

The work of poets and playwrights can be shown to involve an implicit commitment to the hypothesis that minds are (at least) information processing engines. Besides loving, many other familiar states and processes such as seeing, deciding, wondering whether, hoping, regretting, enjoying, disliking, learning, planning and acting all involve various sorts of information processing.

By analysing the requirements for such processes to occur, and relating them to our evolutionary history and what is known about animal brains, and comparing this with what is being learnt from work on artificial minds in artificial intelligence, we can begin to formulate new and deeper theories about how minds work, including how we come to think about qualia, many forms of learning and development, and results of brain damage or abnormality.

But there is much prejudice that gets in the way of such theorising, and also much misunderstanding because people construe notions of "information processing" too narrowly.

1999-2002
Filename: SlomanLogan.toyota.pdf

Title: Evolvable architectures for human-like minds
Authors: Aaron Sloman and Brian Logan
Invited talk at 13th Toyota Conference, on "Affective Minds" Nagoya Japan, Nov-Dec 1999
Published in *Affective Minds,* Ed. Giyoo Hatano, Elsevier, October 2000
Abstract:
There are many approaches to the study of mind, and much ambiguity in the use of words like 'emotion' and 'consciousness'. This paper adopts the design stance, in an attempt to understand human minds as information processing virtual machines with a complex multi-level architecture whose components evolved at different times and perform different sorts of functions. A multi-disciplinary perspective combining ideas from engineering as well as several sciences helps to constrain the proposed architecture. Variations in the architecture should accommodate infants and adults, normal and pathological cases, and also animals. An analysis of states and processes that each architecture supports provides a new framework for systematically generating concepts of various kinds of mental phenomena. This framework can be used to refine and extend familiar concepts of mind, providing a new, richer, more precise theory-based collection of concepts. Within this unifying framework we hope to explain the diversity of definitions and theories and move towards deeper explanatory theories and more powerful and realistic artificial models, for use in many applications, including education and entertainment.

2003

- **http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-cogaff-03.pdf**
Title: Progress report on the Cognition and Affect project:
Architectures, Architecture-Schemas, And The New Science of Mind
(Original 2003. Revised October 2004 and 2008.)
Author: Aaron Sloman
Date installed: 7 Dec 2003 (Revised Oct 2004. Liable to be further revised.)

Abstract:
The 'Cognition and Affect' project, which was called 'The attention and affect' project for a few years (circa 1991-1993), is a continuation of research on the nature of mind in natural and artificial systems by A.Sloman, which began around 1970 while he was at Sussex University, accelerated by a one-year visiting fellowship at the University of Edinburgh in 1972-3, continued during the build up at Sussex of COGS (The School of Cognitive and Computing Sciences) and accelerated further after he moved to the University of Birmingham in 1991.

The work is a mixture of philosophy, science and engineering, concerned especially with the role of explanatory architectures. In this it overlaps with Marvin Minsky's work on The Emotion Machine (2006)

This report was triggered partly by a consultation for DARPA regarding cognitive systems and partly by the need to write a final report for the Leverhulme-funded project on *Evolvable virtual information processing architectures for human-like minds* (1999--2003) on which there were three research fellows in sequence, Brian Logan, Matthias Scheutz and Ron Chrisley. Several PhD students at the University of Birmingham also contributed.

The Leverhulme project has ended but work arising out of it continues, as will the Cognition and Affect project, with or without funding. Ongoing activities include a grand challenge proposal and European Community research initiatives, including this initiative on models of consciousness.

A major new robotic project funded by the EC started in September 2004
CoSy: Cognitive systems for cognitive assistants

More recent changes:
http://www.cs.bham.ac.uk/research/projects/cogaff/#overview
New architecture diagram, showing overlaps between input, output, and "central" components of the architecture:
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/vm-functionalism.html
(also PDF)

2004

- **AAAI 2004 Workshop invited paper:** What are emotion theories about?
Author: Aaron Sloman
Paper:
http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-aaai04-emotions.pdf
Presentation:
http://www.cs.bham.ac.uk/research/projects/cogaff/talks/sloman-aaai04-slides.pdf
Invited talk at cross-disciplinary workshop on "Architectures for Modeling Emotion"
AAAI Spring Symposium at Stanford University in March 2004.
https://aaai.org/Library/Symposia/Spring/ss04-02.php
Abstract:
This is a set of notes relating to a talk given at the cross-disciplinary workshop on Architectures for Modeling Emotion at the AAAI Spring Symposium at Stanford University in March 2004. The organisers of the workshop note that work on emotions "is often carried out in an ad hoc manner", and hope to remedy this by focusing on two themes (a) validation of emotion models and

architectures, and (b) relevance of recent findings from affective neuroscience research. I shall focus mainly on (a), but in a manner which, I hope is relevant to (b), by addressing the need for conceptual clarification to remove, or at least reduce, the ad-hocery, both in modelling and in empirical research. In particular I try to show how a design-based approach can provide an improved conceptual framework and sharpen empirical questions relating to the study of mind and brain. From this standpoint it turns out that what are normally called emotions are a somewhat fuzzy subset of a larger class of states and processes that can arise out of interactions between different mechanisms in an architecture. What exactly the architecture is will determine both the larger class and the subset, since different architectures support different classes of states and processes. In order to develop the design-based approach we need a good ontology for characterising varieties of architectures and the states and processes that can occur in them. At present this too is often a matter of much ad-hocery. We propose steps toward a remedy.

- **2004**
  Filename:
  petters-aaai-ss-04.pdf
  Author: Dean Petters,
  Title Simulating Infant-Carer Relationship Dynamics
  2004 AAAI Spring Symposium
  Eds Eva Hudlicka, and Lola Caqamero, (Program Chairs)
  Abstract

  > Advances in autonomous agent technology have resulted in the potential for implementations of multiple agents to act as psychological theories of complex social and affective phenomena. Simulating attachment behaviours in infancy provides a relatively simple starting point for this type of theory development. The presence of neurophysiological, psychological and other types of data facilitates the validation of architectural theories by constraining these architectures at multiple levels. A seven part design process is described which details how requirements are specified and how design, implementation and evaluation processes are carried out. Two competing theories are proposed, one that involves some deliberation and one that is reactive only.

- **2004**
  Author: Matthias Scheutz
  Title: How to Determine the Utility of Emotions
  Paper: https://aaai.org/Papers/Symposia/Spring/2004/SS-04-02/SS04-02-023.pdf
  2004 AAAI Spring Symposium
  Eds Eva Hudlicka, and Lola Caqamero, (Program Chairs)
  Abstract:
  In this paper, we describe a new methodology for determining the utility of emotions. After briefly reviewing the status quo of emotional agents in AI, we describe the methodology and demonstrate it by showing the utility of "anger" for biologically plausible foraging agents in an evolutionary setting.

- **2005** Invited book chapter:
  Authors: Aaron Sloman, Ron Chrisley and Matthias Scheutz

Title: The Architectural Basis of Affective States and Processes
http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-chrisley-scheutz-emotions-long.pdf
(Original longer version, 2 Dec 2003. Shortened later at request of publisher and editors.)
http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-chrisley-scheutz-emotions.pdf
(truncated for book chapter)

> In *Who Needs Emotions?: The Brain Meets the Robot,*
> Eds. M. Arbib and J-M. Fellous, Oxford University Press, Oxford, New York, 2005, pp. 203--244

The version published by OUP was gratuitously changed and mangled in many ways by the copy-editor who mostly did not understand what was being said, despite strong protestations by the authors. And OUP (at least in the USA) do not understand requirements for publishing scientific texts -- they use out of date style guides designed for literary texts. See http://www.cs.bham.ac.uk/~axs/publishing.html

Abstract:

Much discussion of emotions and related topics is riddled with confusion because different authors use the key expressions with different meanings. Some confuse the concept of "emotion" with the more general concept of "affect", which covers other things besides emotions, including moods, attitudes, desires, preferences, intentions, dislikes, etc. Moreover researchers have different goals: some are concerned with understanding natural phenomena, while others are more concerned with producing useful artefacts, e.g. synthetic entertainment agents, sympathetic machine interfaces, and the like. We address this confusion by showing how "architecture-based" concepts can extend and refine our pre-theoretical concepts in ways that make them more useful both for expressing scientific questions and theories, and for specifying engineering objectives. An implication is that different information-processing architectures support different classes of emotions, different classes of consciousness, different varieties of perception, and so on. We start with high level concepts applicable to a wide variety of types of natural and artificial systems, including very simple organisms, namely concepts such as "need", "function", "information-user", "affect", "information-processing architecture". For more complex architectures, we offer the CogAff schema as a generic framework which distinguishes types of components that may be in a architecture, operating concurrently with different functional roles. We also sketch H-Cogaff, a richly-featured special case of CogAff, conjectured as a type of architecture that can explain or replicate human mental phenomena. We show how the concepts that are definable in terms of such architectures can clarify and enrich research on human emotions. If successful for the purposes of science and philosophy the architecture is also likely to be useful for engineering purposes, though many engineering goals can be achieved using shallow concepts and shallow theories, e.g., producing "believable" agents for computer entertainments. The more human-like robot emotions will emerge, as they do in humans, from the interactions of many mechanisms serving different purposes, not from a particular, dedicated "emotion mechanism".

> There is a summary of a review by Zack Lynch here.
> "Rather than building on the hype surrounding thinking machines the book provides a superb scientific analysis of the current state of emotions research in animals, humans and man-made systems." .... "While technical in parts, this book is an important contribution to the emerging field of emotional neurotechnology. It is a stimulating book that is well edited and researched. I highly recommend Who Needs Emotions? for

researchers and graduate students across neuroscience and computer science."

- **2005** A. Sloman and R. L. Chrisley, 2005, More things than are dreamt of in your biology: Information-processing in biologically-inspired robots, in *Cognitive Systems Research* 6, 2, June, pp. 145--174,
  http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-chrisley-cogsys.pdf

- **2005** Requirements for a Fully Deliberative Architecture
  (Or component of an architecture)
  Aaron Sloman
  http://www.cs.bham.ac.uk/research/projects/cogaff/misc/fully-deliberative.html
  http://www.cs.bham.ac.uk/research/projects/cogaff/misc/fully-deliberative.pdf
  (Originally installed: 2006, after a conversation with Dean Petters, then updated several times.)

- **2009**
  Title: Architecture-Based Motivation vs Reward-Based Motivation
  Author: Aaron Sloman **Local (updated, expanded) version of previously published paper:**
  http://www.cs.bham.ac.uk/research/projects/cogaff/misc/architecture-based-motivation.html
  http://www.cs.bham.ac.uk/research/projects/cogaff/misc/architecture-based-motivation.pdf
  The original version (2009) was an invited article for *Philosophy newsletter:*
  PDF version of newsletter on APA website
  Date Installed: 10 Nov 2009 (Modified: 24 Jan 2014; 14 Jun 2015; ...)

  Where published:

  Originally published as an invited paper in:
  Newsletter on Philosophy and Computers, American Philosophical Association,
  (Including Newsletter index)
  This paper was published in issue 09, 1, pp. 10--13:
  (PDF) version of whole newsletter
  www.apaonline.org/resource/collection/EADE8D52-8D02-4136-9A2A-729368501E43/v09n1Computers.pdf
  (Now partly out of date: see later local version, above.)

  Abstract:

  "Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them." David Hume, A Treatise of Human Nature (2.3.3.4), 1739-1740 (http://www.class.uidaho.edu/mickelsen/ToC/hume%20treatise%20ToC.htm)

  Whatever Hume may have meant by this, and whatever various commentators may have taken him to mean, I claim that there is at least one interpretation in which this statement is obviously true, namely: no matter what factual information an animal or machine A contains, and no matter what competences A has regarding abilities to reason, to plan, to predict, or to explain, A will not actually *do* anything unless it has, in addition, some sort of control mechanism that selects among the many alternative processes that A’s information and competences can support.

In short: *control* mechanisms are required in addition to *factual information and reasoning* mechanisms if A is to do anything. This paper is about what forms of control are possible and their relative merits. In at least some cases there are pre-existing motives, and the control arises out of selection of a motive for action. That raises the question where motives come from. My answer is that they can be generated and selected in different ways, but one way is not itself motivated: it merely involves the operation of mechanisms, based on evolutionary heritage, in the architecture of A, that generate motives and select some of them for action. They could be described as generated by "motivational reflexes" that are directly or indirectly products of our evolutionary heritage. They are valuable because their activation leads to acquisition of information even when there is no need for that information at the time, and the individual has no reason to believe the information will be useful at some time in the future.

The view I wish to oppose is that all motives must somehow serve the *current* interests of A, or be *rewarding* for A. This view is widely held and is based on a lack of imagination about possible designs for working systems. I summarize it as the assumption that all motivation must be reward-based. In contrast, I claim that at least some motivation may be architecture-based, in the sense explained in the paper.
The architecture-based motivation theory is consistent with the theory of the "Meta-configured Genome", developed in collaboration with Jackie Chappell, Biosciences, University of Birmingham, and summarised here:
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/meta-configured-genome.html

---

- **2013**
  Virtual Machine Functionalism (VMF)
  (The only form of functionalism worth taking seriously
  in Philosophy of Mind and theories of Consciousness)
  Author: Aaron Sloman
  http://www.cs.bham.ac.uk/research/projects/cogaff/misc/vm-functionalism.html
  http://www.cs.bham.ac.uk/research/projects/cogaff/misc/vm-functionalism.pdf
  Installed 2013, but updated several times since then.
  Abstract
  Most philosophers appear not to have distinguished the broad concept of *Virtual Machine Functionalism* (VMF), described in Sloman (1993), (2002) and Sloman&Chrisley (2003), from the better known, more restricted, version referred to in those papers as *Atomic State Functionalism* (ASF), which is often erroneously given as an explanation of what "Functionalism" refers to, e.g. in Block (1995). (I don’t think that paper expresses his current views.)

  One of the main differences is that **ASF** encourages talk of supervenience of *states* and *properties*, whereas **VMF** requires supervenience of *machines* that are arbitrarily complex networks of causally interacting (virtual, but real) processes, possibly operating on different time-scales, and not necessarily synchronised with one another -- especially if different substates interact with different parts of the environment that are not synchronised with one another, for example when you are watching waves breaking over rocks while having a conversation with a friend, and walking along an uneven path guided by a handrail, slightly irritated by a stone in your shoe.

Examples include many different processes running concurrently on modern (single-CPU or multi-CPU) computers performing various tasks concerned with handling interfaces to physical devices, managing file systems, handling user-access, dealing with security, providing tools, entertainments, and games, and processing research data.

A less obvious example of *virtual machine functionalism* is the kind of functionalism involved in a large collection of possibly changing socio-economic structures and processes interacting in a complex community. Yet another example is illustrated by the complex network of mental virtual machines involved in the many levels and types of information about spatial structures, processes, and relationships (including percepts of moving shadows, reflections, highlights, optical-flow patterns and changing affordances) processed in parallel as you walk through a crowded car-park on a sunny day: generating a whole zoo of interacting qualia. (Forget solitary red patches, or experiences thereof.)

Keywords: asynchronous concurrent causation, atomic state functionalism, counterfactual conditionals, definability, information processing, interfaces to environment, operating system, physics, qualia, representation, self monitoring, virtual machine functionalism, virtual machine supervenience,

---

ADDITIONAL ITEMS TO BE INSERTED HERE

2014

- Talk 28: (Given on several occasions, with variations)
  Do Intelligent Machines, Natural or Artificial, Really Need Emotions?
  Slides:
  http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#emotions
  Last revised: 14 Jan 2014
  Aaron Sloman
  Abstract

Since the publication of the book *Descartes' Error* in 1994 by Antonio Damasio, a well-known neuroscientist, it has become very fashionable to claim that emotions are necessary for intelligence. I think the claim is confused and the arguments presented for it fallacious.

Part of the problem is that many of the words we use for describing human mental states and processes (including 'emotion' and 'intelligence') are far too ill-defined to be useful in scientific theories. Nevertheless there are many people who LIKE the idea that emotions, often thought of as inherently irrational, are required for higher forms of intelligence, the suggestion being that rationality is not all it's cracked up to be. But wishful thinking is not a good basis for advancing scientific understanding.

> Another manifestation of wishful thinking is people attributing to me opinions that are the opposite of what I have written in things they claim to have read.

So I propose that we investigate, in a dispassionate way, the variety of design options for minds, whether in animals (including humans) or machines, and try to understand the trade-offs between different ways of assembling systems that survive in a complex and

changing environment. This can lead to a new science of mind in which the rough-hewn concepts of ordinary language (including garden-gate gossip and poetry) are shown not to be wrong or useless, but merely stepping stones to a richer, deeper, collection of ways of thinking about what sorts of machines we are, and might be.

2017

- **Architectures underlying cognition and affect in natural and artificial systems**
  http://www.cs.bham.ac.uk/research/projects/cogaff/aisb17-emotions-sloman.html
  http://www.cs.bham.ac.uk/research/projects/cogaff/aisb17-emotions-sloman.pdf
  Extended Abstract for invited talk at AISB 2017
  Aaron Sloman

  This is a summary of some of the ideas in my invited talk for the Symposium on "Computational modelling of emotion: theory and applications" at AISB 2017. A deep understanding of human (or animal) minds requires a broad and deep understanding of the types of information processing functions and information processing mechanisms produced by biological evolution, and how those functions and mechanisms are combined in architectures of increasing sophistication and complexity over evolutionary trajectories leading to new species, and how various kinds of evolved potential are realised by context-sensitive mechanisms during individual development. Some aspects of individual development add context-specific detail to products of the evolutionary history, partly because evolution cannot produce pre-packaged specifications for complete information processing architectures, except for the very simplest organisms. Instead, for more complex organisms, including humans, different architectural layers develop at different times during an individual's life, partly under the influence of the genome and partly under the influence of what the individual has so far experienced, learnt, and developed. This is particularly obvious in language development in humans, but that is a special case of a general biological pattern (identified in joint work with Jackie Chappell, partly inspired by theories of Annette Karmiloff-Smith, among others). This paper complements a paper presented in the Symposium on Computing and Philosophy at AISB 2017, which develops more general ideas about evolution of information processing functions and mechanisms, partly inspired by Turing's work on morphogenesis:
  http://www.cs.bham.ac.uk/research/projects/cogaff/sloman-aisb17-CandP.pdf

---

- **Cognition and Affect: Past and Future**
  Cognition and Affect Workshop
  Following AISB 2017 Discussions http://aisb2017.cs.bath.ac.uk/
  Monday 24th April 2017
  School of Computer Science
  University of Birmingham, UK
  http://www.cs.bham.ac.uk/research/projects/cogaff/misc/cogaff-sem-apr-2017.html

  The workshop summarised here was held on Monday afternoon 24th April 2017. People who had attended or communicated about the workshop were invited to submit comments related to the discussion at the workshop or the document on **anger** mentioned in the workshop "homework":
  http://www.cs.bham.ac.uk/research/projects/cogaff/misc/cogaff-sem-apr-2017.html#homework

---

- **An example of conceptual analysis:**
  What does 'X is angry with Y' mean? Aaron Sloman, April 2017
  Notes prepared for the above workshop after asking intending participants to attempt an analysis of the concept of "anger".
  Incomplete sample analysis available here:
  http://www.cs.bham.ac.uk/research/projects/cogaff/misc/anger.html
  (also pdf).

REFERENCES AND LINKS

-
-
-
-

A partial index of discussion notes in this directory is in
http://www.cs.bham.ac.uk/research/projects/cogaff/misc/AREADME.html

---

Maintained by Aaron Sloman
School of Computer Science
The University of Birmingham